Gerald Amiel Ballena

gmballena@up.edu.ph github.com/GABallena linkedin.com/in/gerald-amiel-ballena
Metro Manila, Philippines orcid.org/0009-0000-8857-9755

Professional Summary

Versatile bioinformatics specialist with focused on general expertise in scalable workflows, computational biology, and machine learning. Skilled in developing scalable pipelines and implementing robust methodologies to tackle complex biological problems.

Experience

2024 – Present **Project Technical Specialist**
University of the Philippines, College of Public Health

- Developed scalable bioinformatics workflows for high-throughput sequence analysis.
- Applied metagenomic profiling techniques to environmental datasets.
- Collaborated with cross-disciplinary teams on projects involving data visualization and computational modeling.

Skills

Programming

Languages: Python, R, Bash, Perl

Installation Toolkits: BiocManager, upcxx

Bioinformatics Tools

Metagenomics: Kraken2, MetaPhlAn, HUMAnN

Assembly & Binning: MEGAHIT, METAWRAP, CheckM

Genomics: CLC Workbench, Roary, MinHash

Quality Control: FastQC, BUSCO, QUAST **Trimming**: Trimmomatic, Sickle, Cutadapt

Phylogenetics: RaxML, FastTree, IQ-TREE, phyML, BEAST

Gene Ontology: KEGG, GOseq

Annotation: Prokka, ShortBRED, EggNOG-mapper

Others: DESeq2, BBMap Suite, anvio-8, SPAdes

Data Analysis & Visualization

Machine Learning: scikit-learn, caret (R)

Visualization: ggplot2, Plotly, Krona, Shiny, Seaborn, Tableau

Technical Skills

Version Control: Git, GitHub

High-Performance Computing: Slurm

Workflow Automation: Snakemake, Conda, YAML (config files)

Virtualization: Docker, WSL2

Other Tools

Technical Writing: Overleaf, TeXStudio

References available only upon request and if shortlisted

Collaboration: Thunderbird, Notion, MS Word (mailing lists)

Documentation: TeXStudio, Jupyter Notebooks

Certifications and Relevant Coursework

Certifications

AI Fundamentals: Nov 2024

Data Literacy: Nov 2024

Fundamentals Courses

Introduction to Data Engineering: Nov 2024

Intermediate-Level Courses

Intermediate R: Nov 2024

Introduction to Bioconductor in R: Nov 2024

RNA-Seq with Bioconductor in R: Nov 2024

Differential Expression Analysis with limma in R: Nov 2024

ChIP-Seq with Bioconductor in R: Nov 2024

Analyzing Genomic Data in R: Nov 2024

Advanced-Level Courses

Hyperparameter Tuning in R: Nov 2024

Designing Machine Learning Workflows in Python: Nov 2024

Ongoing Courses

Ensemble Methods in Python: Bagging, Boosting, Stacking

Visualizing Geospatial Data in R: Visualizing complex spatial datasets for actionable insights.

Assessments

Azure Fundamentals: Advanced | Score: 180 | Percentile: 99th

Data Storytelling: Advanced | Score: 200 | Percentile: 99th

Projects

For more details on public projects, visit GitHub Documentation.pdf.
Note Some repositories including Documentation repo access is currently restricted due to NDA obligations.

Scripts Repositories

Side repo: Scripts for general-purpose tasks, including package management and workflow setup.

Experimental Repository: Exploratory scripts focused on innovative computational techniques.

Health repo: Tools and scripts aimed at health-related data analysis and personal wellness tracking.

Finance repo: **WIP**; developing scripts for personal finance management.

Project4 repo: Scripts designed for NGS analyses and workflow automation; details kept private due to project confidentiality.

Kitchen repo: We cookin' some unconventional ideas here.

Documentation Repositories

Documentation repo: Explanation of every script I've ever written; written in a non-technical tone for non-technical audiences.

Confidential repo: LaTeX-compiled reports and documentation of my work as PTS I; restricted access due to confidentiality.

Hunting repo: Includes this CV, my resume, associated `.tex` files, and certifications; tailored for job hunting.

References available only upon request and if shortlisted

Key Projects

Bioinformatics Workflow Development: Designed and implemented modular pipelines for omics data analysis, emphasizing quality control, data assembly, and scalable workflows. Focused on creating reproducible methodologies applicable across diverse datasets.

Spatial Data Analysis: Utilized QGIS, R and geospatial libraries to visualize spatial patterns in biological datasets. Developed custom visualization tools for integrating geospatial and genomic data in ecosystem-level studies.

Public Health Data Exploration: Developed computational approaches for microbial profiling and identifying markers of interest in public health datasets. Leveraged metagenomic techniques to enable actionable insights for research studies.

Development of K-mer Analysis and Statistical Workflow:

- **Objective:** Build an advanced pipeline for k-mer generation, variance analysis, and distribution modeling.
- **Details:** Focused on creating statistical workflows to evaluate parametric and non-parametric fits, automating evaluations with custom Python scripts.
- **Tools & Technologies:** Jellyfish, Python, Snakemake, Conda.

High-Accuracy Sequence Alignment Workflow:

- **Objective:** Create a modular workflow for high-identity sequence alignment with validation across multiple tools.
- **Details:** Integrated tools like Bowtie2, BWA, and Minimap2 with standardized outputs to enable comparative analysis.
- **Tools & Technologies:** Bowtie2, BWA, Minimap2, KMA, Python, Snakemake.

Shannon Entropy Analysis of K-mers and Taxonomic Profiles:

- **Objective:** Quantify sequence complexity and diversity using entropy-based metrics.
- **Details:** Developed workflows to calculate and integrate Shannon entropy metrics for assessing genomic diversity and data quality.
- **Tools & Technologies:** Jellyfish, Kraken2, Python, Snakemake.

Comprehensive Workflow for Contaminant Removal and SCG Validation:

- **Objective:** Design a robust pipeline for contaminant filtering and SCG validation.
- **Details:** Automated filtering using k-mer-based methods and validated results with SCG retention analysis via BUSCO.
- **Tools & Technologies:** BUSCO, Jellyfish, Python, Snakemake, Conda.

Pipeline Automation and Dependency Management:

- **Objective:** Streamline bioinformatics workflows with automated dependency checks and alias creation.
- **Details:** Developed:
  - `bioconda_search.py` to identify relevant bioinformatics tools.
  - `create_aliases.py` to automate Bash alias generation.
  - `check_dependencies.py` and `append_new.py` for managing and updating YAML files.
- **Tools & Technologies:** Python, YAML, Conda, Bioconda.

Version Management and Workflow Validation:

- **Objective:** Ensure reproducibility in bioinformatics workflows.
- **Details:** `versioncheck.bash` validates installed tool versions against expected configurations, improving workflow consistency.
- **Tools & Technologies:** Bash, Python, YAML.

References available only upon request and if shortlisted

Ongoing Coursework

Ensemble Methods in Python Visualizing Geospatial Data in R

Languages

English: Fluent — *Professional proficiency.*

Tagalog: Native — *Conversational*

Ilocano: Can understand — *Wernicke's area skill issue.*

Computer Languages (For Fun)

Python: Fluent — *Fluent enough for advanced bioinformatics.*

R: Advanced — *Best data visualization tool out there (currently).*

Bash: Intermediate — *Go-to shell scripting language.*

Perl: Fair — *Would rather "speak" in Bash.*

C++: Basic — *As fluent as someone who hasn't used it in a decade.*

References

Available upon request & only if already shortlisted.